iWARP Brings Low-Latency Fabric Technology to Ethernet

Authors:

David Fair, Intel

Manoj Wadekar, QLogic

Blaine Kohl, Ethernet Alliance

1. Introduction

In November 2009, the Ethernet Alliance announced a focused group effort inside the Ethernet in the Data Center subcommittee to support market development for the iWARP technology. As a technology, iWARP is becoming mature with multiple vendors developing and shipping iWARP adapters – known as rNICs – and with production operating system support from Linux and Microsoft. The University of New Hampshire Interoperability Lab (UNH-IOL) has established an iWARP Consortium to assist vendors in testing for iWARP compliance and interoperability.

The Ethernet Alliance support comes at a time when iWARP is starting to gain momentum in the industry as a means to provide a low-latency, high performance computing (HPC) solution. Several large HPC installations are successfully using iWARP including NYSE Euronext financial services. With the addition of iWARP to the Ethernet in the Data Center subcommittee, there is now the capability to showcase Ethernet as *THE* converged network with the ability to support LAN traffic, storage traffic and interprocessor communication.

2. iWARP Overview

iWARP (internet wide-area RDMA protocol) is an Ethernet technology that permits low-latency transmission over IP using TCP/IP, or streaming control transport protocol (SCTP) transports. iWARP brings low-latency fabric technologies to data centers committed to Ethernet. It works with the existing infrastructure of Ethernet switches and routers. The specification was developed by the RDMA Consortium and the standard is maintained by the Internet Engineering Task Force (IETF).

iWARP shares verbs with InfiniBand. The iWARP software stack was developed and is maintained through the open-source efforts of the Open Fabric Alliance. The iWARP stack is already deployed in production Linux and an iWARP stack for Windows Server 2008 is available from Microsoft (Network Direct). Several vendors have RDMA NICS (rNICs) supporting iWARP in production. UNH-IOL has established an iWARP Consortium focused on iWARP compliance and interoperability testing.

With the rapid deployment ramp of 10 Gigabit Ethernet (10 GbE) in the data center underway and the emergence of cost-effective, dense 10 GbE switches, iWARP is poised for significant penetration of the data center. However, market awareness of iWARP and its benefits among IT managers is currently weak. For this reason a group of industry stakeholders in iWARP are forming an “iWARP” group for the promotion and advancement of iWARP within the Ethernet Alliance. This work will take place within the Ethernet in the Data Center subcommittee, which is chartered to be a reference and resource for IT professionals for both existing and emerging data center-focused Ethernet technologies and standards.

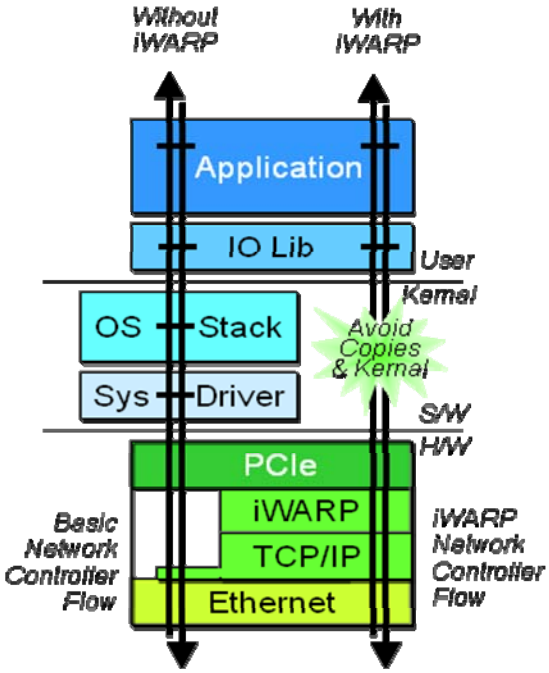
The heart of what iWARP brings to Ethernet is Remote Direct Memory Access (RDMA) capability. RDMA is the capability to write data directly from the memory of one computer into the memory of another with minimal operating system engagement. RDMA enables very low-latency data transmissions. RDMA can enable zero-copy data transmission and does in iWARP via a protocol called Direct Data Placement Protocol (DDP). Note, however, in iWARP the transmission is performed by the underlying industry-standard transport TCP. Application read and write requests are handled by the RDMA Protocol (RDMAP) which establishes a connection between the two computers and passes the request to direct data placement (DDP).

Using TCP as iWARP's transport does create a challenge, since TCP is not based on transmitting distinct messages also known as protocol data units (PDUs) but rather just on sequences of bytes. Framing to guarantee message boundaries is accomplished with marker PDU alignment (MPA).

The iWARP stack is known as the IETF Remote Direct Data Placement (RDDP) group and consists of these three protocols – RDMAP, DDP, and MPA. The iWARP stack may sound complicated, but it is all implemented in the silicon of an iWARP rNIC and thus invisible to the user.

iWARP delivers improved performance by:

- **Eliminating Intermediate Buffer Copies:** Data is placed directly in application buffers vs. being copied multiple times to driver and network stack buffers, thus freeing up memory bandwidth and CPU compute cycles for the application.
- **Delivering a Kernel-Bypass Solution:** Placing data directly in user space avoids kernel-to-user context switches which adds additional latency and consumes additional CPU cycles that could otherwise be used for application processing.
- **Accelerated TCP/IP (Transport) Processing:** TCP/IP processing is done in silicon/hardware versus operating system network stack software, thereby freeing up valuable CPU cycles for application compute processing. The hardware that accomplishes this acceleration is known as a TCP Offload Engine (TOE) and is deployed more broadly that just for iWARP. However, TOE is critical for the performance of iWARP.



iWARP gains significant advantages for being implemented on top of industry standard network and transport protocols. Perhaps most importantly, iWARP can use existing and familiar data

center management tools and methods. iWARP runs over IP. Thus, iWARP traffic is routable across subnets enabling scalability that other RDMA technologies lack, including InfiniBand. IP extensions (e.g., IPSEC) are immediately available to iWARP. Because iWARP runs over the TCP protocol, it interoperates with all data center equipment (switches, security appliances, etc.)

Bringing RDMA to Ethernet, iWARP lends itself to environments that require low-latency performance in an Ethernet ecosystem, including HPC (High Performance Computing) Clusters, Financial Services, Enterprise Data Centers, and Clouds. All of which value Ethernet as an existing, reliable, and proven IT environment that uses heterogeneous equipment and widely-deployed management tools.

3. About Ethernet Alliance

The Ethernet Alliance is a community of Ethernet end users, system and component vendors, industry experts and university and government professionals who are committed to the continued success and expansion of Ethernet. The Ethernet Alliance brings Ethernet standards to life by supporting activities that span from incubation of new Ethernet technologies to interoperability demonstrations, certification and education.